AutoML e Redução de Dimensionalidade: Análise Comparativa Entre Modelos de Aprendizado de Máquina

AutoML and Dimensionality Reduction: Comparative Analysis Among Machine Learning Models

Alvaro Henrique Nunes de Lima[1], Danilo Sipoli Sanches[2]

RESUMO

O Aprendizado de Máquina Automatizado (AutoML) busca simplificar e acelerar o ciclo de desenvolvimento de modelos preditivos, com objetivo de facilitar o acesso a modelos de predição mesmo com pouco conhecimento na área. Este estudo apresenta uma análise entre abordagens de AutoML comparado a um modelo tradicional (Support Vector Machine) em conjuntos de dados distintos. O AutoML foi implementado por meio das bibliotecas AutoSklearn e FlaML, e avaliados utilizando a acurácia. Além disso, para reduzir ainda mais o tempo e eficiencia, uma redução de dimensionalidade foi aplicada com base nas características mais relevantes com a ajuda do método SHAP (SHapley Additive exPlanations). Os resultados destacam que o AutoML geralmente superam o modelo tradicional em termos de acurácia, com desempenho variando dependendo da complexidade do conjunto de dados. Esta pesquisa contribui para a compreensão do potencial do AutoML em agilizar o desenvolvimento de modelos de aprendizado de máquina.
PALAVRAS-CHAVE: Aprendizado de Máquina; AutoML; Redução de Dimensionalidade.

ABSTRACT

Automated Machine Learning (AutoML) aims to simplify and expedite the predictive model development cycle, making predictive models accessible even to those with limited expertise in the field. This study presents a comparative analysis between AutoML approaches and a traditional model (Support Vector Machine) on distinct datasets. AutoML was implemented using the AutoSklearn and FlaML libraries and evaluated using accuracy. Furthermore, to further enhance efficiency, dimensionality reduction was applied based on the most relevant features using the SHAP (SHapley Additive exPlanations) method. The results highlight that AutoML generally outperforms the traditional model in terms of accuracy, with performance varying depending on the dataset's complexity. This research contributes to understanding the potential of AutoML on making the development of machine learning models more efficient.
KEYWORDS: Machine Learning; AutoML; Dimensionality Reduction.

INTRODUÇÃO

Nos últimos anos, a área de Aprendizado de Máquina (ML) tem acompanhado um notável aumento na adoção de algoritmos de automação, os quais simplificam processos anteriormente dependentes de análises e implementações detalhadas. (HE; ZHAO; CHU, 2021). Essa transformação foi impulsionada pelo desenvolvimento de algoritmos de Aprendizado de Máquina Automatizado (AutoML), que buscam simplificar e acelerar o ciclo de desenvolvimento de modelos preditivos (FEURER et al., 2015). O AutoML busca automatizar esse processo, tornando a análise de dados mais

[1] Bolsista do(a) Fundação Araucária. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: alvarolima@alunos.utfpr.edu.br. ID Lattes: 6764199965295743.
[2] Docente no Curso de Engenharia de Computação. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil. E-mail: danilosanches@utfpr.edu.br. ID Lattes: 6377657274398145.

acessível a uma variedade de usuários, independentemente de sua experiência em ML. No entanto, é fundamental entender se essa automatização não apenas economiza tempo, mas também produz modelos que são tão eficazes quanto os desenvolvidos sem automatização.

Além disso, outro fator importante que pode afetar o desempenho da predição é a dimensionalidade do conjunto de dados, especialmente em grandes conjuntos de dados, onde esse tempo pode ser substancial (MÜLLER; GUIDO, 2017). Nesse contexto, é importante empregar técnicas de extração de características, as quais podem ajudar a mitigar esse problema. A abordagem escolhida é o SHAP (SHapley Additive exPlanations), uma técnica baseada na teoria dos jogos cooperativos que calcula a contribuição de cada característica para uma predição do modelo (LUNDBERG; LEE, 2017).

Neste estudo, foi realizada uma análise comparativa da acurácia e redução de dimensionalidade entre duas principais bibliotecas de AutoML, AutoSklearn (FEURER et al., 2015) e FlaML(WANG et al., 2019), além de um modelo *Support Vector Machine* (SVM), complementado por uma otimização de hiperparâmetros. O objetivo é avaliar se os resultados obtidos com abordagens de AutoML apresentam uma melhora significativa em relação ao modelo tradicional.

MATERIAIS E MÉTODOS

A tarefa principal do estudo foi a classificação, binária ou multiclasse, de conjuntos de dados selecionados por meio de aprendizagem supervisionada. Dada a diversidade de técnicas de aprendizado de máquina empregadas na pesquisa, a métrica de acurácia foi escolhida como referência para a comparação dos modelos. Os três conjuntos de dados experimentais foram obtidos a partir do repositório *UCI Machine Learning Repository*. Cada conjunto de dados foi escolhido de forma a apresentar características distintas, com o objetivo de avaliar a eficácia do AutoML em diferentes cenários. O Quadro 1 destaca quais conjuntos foram selecionados, sendo os critérios fundamentais que eles possuíssem menor dimensionalidade e contivessem dados reais destinados à tarefa de classificação.

Quadro 1 – Conjuntos de dados utilizadas

nº	Conjunto de Dados	Características	Instancias	Tipo de Classificação
1	*Parkinson's Disease Prediction* [1]	23	197	Binário
2	*Student Dropout and Academic Success* [2]	36	4424	Multiclasse
3	*DARWIN: Alzheimer Prediction Dataset* [3]	451	174	Binário

Fonte: Elaborado por autores (2022).

Cada conjunto de dados passou por três algoritmos diferentes: SVM/Grid Search, AutoSklearn e FlaML. Após o treinamento, os modelos resultantes foram analisados pelo SHAP para identificar as características mais relevantes em cada caso. A partir dessa lista de características essenciais, aplicou-se uma técnica de redução de dimensionalidade para remover aquelas que tiveram pouco ou nenhum impacto na predição. Os testes realizados foram comparados usando a métrica de acurácia

[1] Disponível em: https://archive.ics.uci.edu/dataset/174/parkinsons
[2] Disponível em: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
[3] Disponível em: https://archive.ics.uci.edu/dataset/732/darwin

e representados visualmente por meio de uma matriz de confusão, que facilita a identificação de erros e acertos dos modelos.

BIBLIOTECAS DE DESENVOLVIMENTO

O código foi desenvolvido em Python utilizando a Scikit-learn como principal biblioteca de implementação dos modelo e métodos empregados sobre o SVM. Essa biblioteca oferece métodos de pré-processamento de dados, seleção de modelos e avaliação (MÜLLER; GUIDO, 2017).

O AutoSklearn é uma biblioteca de AutoML em Python que se baseia no scikit-learn, no qual realiza a busca pelo melhor modelo e hiperparâmetros utilizando otimização bayesiana (FEURER et al., 2015). Outra importante biblioteca de AutoML é o FlaML, no qual se destaca por sua eficiência na otimização de pipelines de aprendizado de máquina (WANG et al., 2019). Importante ressaltar que ambos os algoritmos de AutoML funcionam durante um tempo determinado pelo programador, logo cada conjunto de dados teve seu tempo proporcional para finalizar o processo.

O *SHAP* é uma biblioteca que atribui a cada característica da base um valor de contribuição para a predição do modelo, baseado em valores de *Shapley* da teoria dos jogos cooperativos (LUNDBERG; LEE, 2017)

RESULTADOS

Nesta seção, apresentamos os resultados obtidos a partir das análises realizadas pelos modelos de AutoML. Os resultados englobam a performance desses modelos nos conjuntos do Quadro 1, abrangendo desde a acurácia até a redução de dimensionalidade. As Tabelas 1, 2 e 3 apresentam os resultados obtidos para cada conjunto e as Figuras 1, 2 e 3 apresentam as matrizes confusão com as quantidades de acertos e erros.

PARKINSONS'S DISEASE PREDICTION

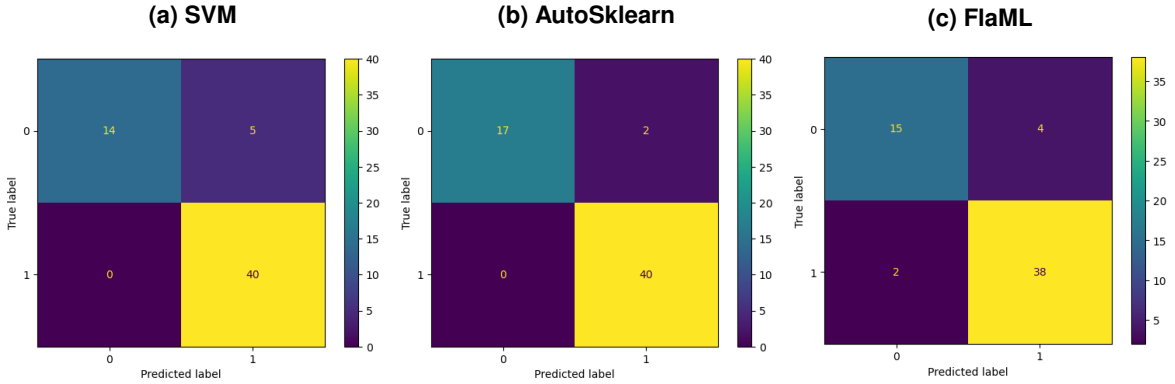
Esse conjunto de dados apresenta valores reais de frequência de voz e diversos outros parâmetros em pacientes com ou sem presença da doença de parkinson. Foi utilizado 70% treino para os modelos de SVM, AutoSklearn e FlaML, e 50 amostras aleatórias para computar as características mais importantes pelo SHAP. Além disso, para este conjunto o AutoSklearn escolheu um algoritmo KNN (*K-Nearest Neighbors*), enquanto o FlaML escolheu um algoritmo LGBM (*Light Gradient Boosting Machine*).

Tabela 1 – Resultados para o conjunto 1

Classificador	Acurácia %	Redução %
SVM	91,525	55
AutoSklearn	96,610	68
FlaML	89,831	86

Fonte: Elaborado pelos autores (2023).

Figura 1 – Matrizes confusão dos modelos para a base Parkinson



Fonte: Elaborado pelos autores (2023).

STUDENT DROPOUT AND ACADEMIC SUCCESS

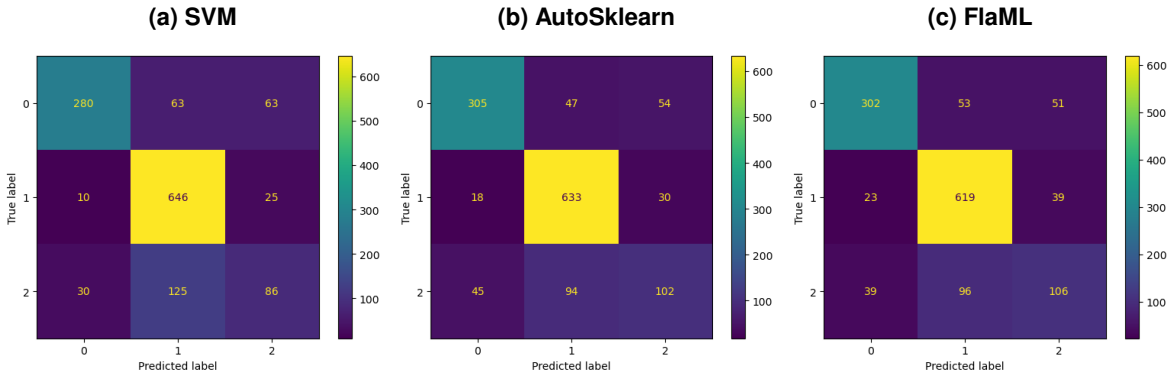
Esse conjunto de dados apresenta dados de estudantes universitários, com a tarefa de classificar caso um estudante se graduou, saiu ou permanece na instituição. Foi utilizado 70% treino para os modelos de SVM, AutoSklearn e FlaML, e 50 amostras aleatórias para computar as características mais importantes pelo SHAP. Nesse teste, o AutoSklearn escolheu um algoritmo *Random Forest* enquanto o FlaML escolheu um algoritmo LGBM.

Tabela 2 – Resultados para o conjunto 2

Classificador	Acurácia %	Redução %
SVM	76,204	73
AutoSklearn	78,313	11
FlaML	77,334	65

Fonte: Elaborado pelos autores (2023).

Figura 2 – Matrizes confusão dos modelos para base Student



Fonte: Elaborado pelos autores (2023).

DARWIN

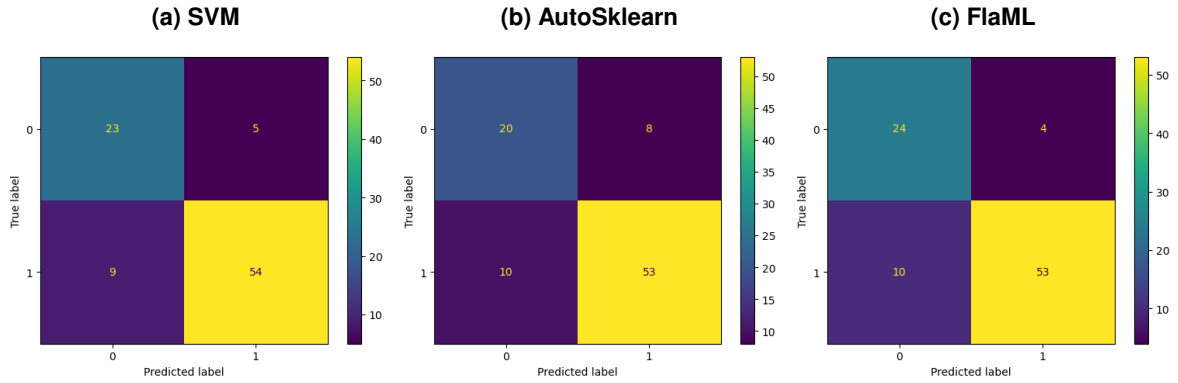
DARWIN é um conjunto de dados com valores reais para classificação de pessoas com a doença do Alzheimer. Foi utilizado 70% da base para treinar os modelos de SVM, AutoSklearn e FlaML, e 50 amostras aleatórias para computar as características mais importantes pelo SHAP. Por fim, nesse teste o AutoSklearn escolheu um algoritmo *Random Forest* enquanto o FlaML escolheu um algoritmo LGBM.

Tabela 3 – Resultados para o conjunto 3

Classificador	Acurácia %	Redução %
SVM	88,6792	62
AutoSklearn	90,5660	42
FlaML	94,3396	78

Fonte: Elaborado pelos autores (2023).

Figura 3 – Matrizes confusão dos modelos para base DARWIN



Fonte: Elaborado pelos autores (2023).

CONCLUSÃO

Em conclusão, os resultados evidenciam que os algoritmos de AutoML geralmente superam o modelo convencional em termos de acurácia. O AutoSklearn se destacou em conjuntos de dados com menos características, enquanto o FlaML demonstrou um desempenho superior apenas no conjunto com grande número de características. Isso sugere que o FlaML utiliza modelos mais complexos, o que o beneficia em conjuntos de dados mais extensos.

Outro ponto relevante é que a redução de dimensionalidade, com base no SHAP, varia consideravelmente dependendo do modelo escolhido. Apesar dessa variação, a maioria dos casos demonstra uma redução significativa, o que contribui para a melhoria do tempo de predição dos modelos.

Agradecimentos

Agradeço a oportunidade dada pelo meu orientador Danilo Sipoli Sanches para realizar esta pesquisa. Também agradeço a UTFPR e Fundação Araucária pelo fomento.

Disponibilidade de Código

O código utilizado para desenvolver as análises está disponível no site do GitHub por meio do link *https://github.com/alfarrh/Pesquisa-AutoML*.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

FEURER, Matthias et al. Efficient and Robust Automated Machine Learning. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 1 (Ed.), 2015, Montréal. **Proceedings...** Cambridge: MIT Press, 2015. v. 28, p. 2755–2763.

HE, Xin; ZHAO, Kaiyong; CHU, Xiaowen. AutoML: A survey of the state-of-the-art. **Knowledge-Based Systems**, Hong Kong, v. 212, n. 1, p. 106622, jan. 2021. ISSN 0950-7051.

LUNDBERG, Scott M; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 1 (Ed.), 2017, Long Beach. **Proceedings...** Seattle: Curran Associates, Inc., 2017. v. 30.

MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to Machine Learning with Python**. 1th. Sebastopol: O'Reilly Media, 2017. ISBN 9781449369415.

WANG, Chi et al. FLAML: A Fast and Lightweight AutoML Library, 1. In: MACHINE LEARNING AND SYSTEMS, 1 (Ed.), 2021. **Proceedings...** [S.l.]: CoRR, 2019. v. 3, p. 434–447.